FINANCE & TECHNOLOGY

Journal of Computational Intelligence in Finance

BAYESIAN YING-YANG DIMENSION REDUCTION AND DETERMINATION

LEI XU

A new general theory is proposed for dimension reduction and determination (DRD), based on the so-called Bayesian Ying-Yang (BYY) learning theory developed in recent years. This theory not only includes conventional factor analysis, principal component analysis (PCA), and nonlinear PCA by least mean squared error reconstruction (LMSER) as special cases, but also provides a unified general guidance for developing various linear and nonlinear DRD techniques and for determining the dimension k of the reduced subspace. As examples, we provide (a) a new algorithm for factor analysis in both batch and adaptive modes, (b) criteria for determining the number of factors and the dimension of the PCA subspace, (c) a procedure for implementing a specific nonlinear BYY DRD based on gaussian mixtures, and (d) extensions for auto-association and LMSER-based nonlinear PCA. Some experimental results are provided.

1 INTRODUCTION

Complexity reduction and *dimensionality reduction* are two commonly used terms describing important strategies that are widely used in data processing and analysis. Usually, the richness of data in a discrete or symbolic representation is measured by complexity, and the mapping from a discrete or symbolic form to another discrete or symbolic form involves complexity reduction. The key point of *Dimension reduction* is mapping the data $x \in \Re^d$ into a lower dimension space y $y \in \Re^k$, $k < d$. The two strategies tackle similar problems from different perspectives. This paper focuses on the second strategy of *dimensionality reduction*.

Specifically, dimension reduction via a linear mapping, $y = Wx$, or a nonlinear mapping, $y = f(x,W)$, is called *Linear* or *Nonlinear Reduction*, respectively, where $f(x,W)$ is a parametric nonlinear function. The purposes of dimension reduction can be roughly divided into two major types. One is to map x into 2 or 3 dimension spaces such that the data structure can be visualized and studied interactively. The other is to preprocess x for the purpose of helping subsequent analyses or pattern recognition. For the first type, we have $k = 2$ or 3 fixed, and the key problem is how to design the mapping such that the data structure in y is kept as close as possible with that in x. For the second type, the problem consists of two important issues. One is *how to determine the dimension of k*, which is called *dimension determination*. The other one is how to design the mapping under a given k. Extensive studies can be found for both linear and nonlinear reduction of both types in the literature on statistical pattern recognition and neural networks over the past few decades. Readers are referred to Samon [1969]; Devijver and Kittler [1982]; Kohonen [1995] for the first type, and to Bourlard and Kamp [1989]; Oja [1983, 1989], and Xu [1993, 1994] for the second type. Tukey [1977] provides insight on how to determine the best approach to conditioning modeling data.

This paper introduces a new general theory for both linear and nonlinear dimension reduction and determination (DRD) based on the so-called Bayesian Ying-Yang (BYY) learning theory developed in recent years [see Xu 1995, 1997a]. This theory not only includes, as special cases, conventional factor analysis, principal components analysis (PCA), and least mean squared error reconstruction (LMSER) based on nonlinear PCA, but also provides a unified general guidance for developing various linear and nonlinear DRD techniques and for determining the dimension k of the reduced subspace, along with several specific case studies.

Similar to PCA and other existing methods mentioned above, the proposed new approach can also be applied to various financial data processing and analyses for data dimension reduction. Roughly speaking, it is applicable to both financial time series $x(t)$ and second hand data, which consists of sets of feature or measurement vectors, $x \in \Re^d$, where each element of x is a feature variable or attribute that is extracted or obtained by some means from the time series. For such second hand data, the approach proposed in this paper can be directly applied. For financial time series $x(t)$, we can perform dimension reduction by decomposing a length-d-segment of time series into $x(t) = \sum_{j=1}^{k} y_j e_j(t)$ with $e_j(t)$, $j=1$, ..., k being a set of bases or eigen-segments. Thus, $y=[y_1, ..., y_k]^T$ is used as a lower dimension representation of the original segment $x(t)$.

2 BYY DIMENSION REDUCTION SYSTEM AND THEORY

2.1 BASIC IDEA OF BAYESIAN YING-YANG LEARNING

The details of Bayesian Ying-Yang (BYY) learning system and theory and its applications can be found in Xu [1997a]. Here, we will only describe its basic concepts.

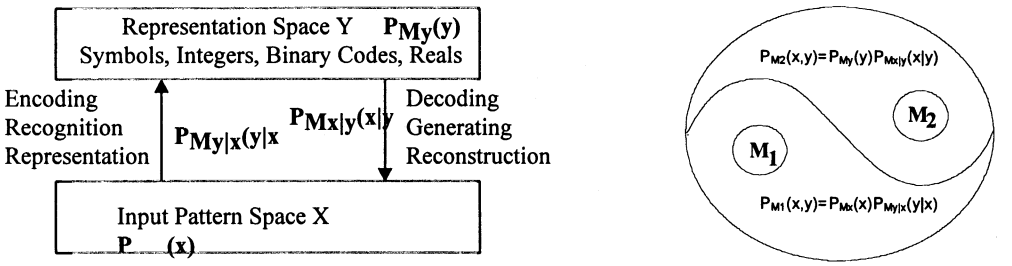


FIGURE 1: The joint input-representation spaces X, Y and the Bayesian YING-YANG system

As shown in Figure 1, the tasks can be summarized as a problem of estimating the joint distribution $p(x, y)$ of the observable pattern x in the observable space X and its representation pattern y in the representation space Y. In the Bayesian framework, we have two complementary representations $p(x, y)=p(y|x)p(x)$ and $p(x, y)=p(x|y)p(y)$. We use two sets of models $M_1 = \{M_{y|x}, M_x\}$ and $M_2 = \{M_{x|y}, M_y\}$ to implement each of the two representations:

$$p_{M_1}(x, y) = p_{M_{y|x}}(y|x)p_{M_x}(x), \quad p_{M_2}(x, y) = p_{M_{x|y}}(x|y)p_{M_y}(y) \tag{1}$$

We call M_x a *Yang* (visible) model, which describes $p(x)$ in the visible domain X, and M_y a *Ying* (invisible) model which describes $p(y)$ in the invisible domain Y. Also, we call the passage $M_{y|x}$ for the flow $x \to y$ a *Yang* (male) passage since it performs the task of transferring a pattern (a real body) into a code (a seed). We call a passage $M_{x|y}$ for the flow $y \to x$ a *Ying* (female) passage since it performs the task of generating a pattern (a real body) from a code (a seed). Together, we have a *YANG* machine M_1 to implement $p_{M_1}(x, y)$ and a *YING* machine M_2 to implement $p_{M_2}(x, y)$. A pair of YING-YANG machines is called a YING-YANG pair or a Bayesian YING-YANG system. This formalization is derived from a famous ancient Chinese philosophy that *every entity in the universe involves the interaction between YING and YANG.*

The task of specifying a Ying-Yang system consists of specifying all the aspects (e.g., the forms of the variables and distributions, the architectures and scales, parameters, etc.) of the four components $p_{M_{y|x}}(y|x)$, $p_{M_x}(x)$, $p_{M_{x|y}}(x|y)$, $p_{M_y}(y)$, which is called *learning* in the broad sense and occurs according to a general principle called *Ying-Yang Harmony*, by minimizing a harmony measure called the *separation function*. Three categories of separation functions have been suggested in Xu [1997a]. One particular example is the minimization of the following Kullback divergence function:

$$KL_{M_1,M_2} = \int p_{M_{y|x}}(y|x)p_{M_x}(x) \ln \frac{p_{M_{y|x}}(y|x)p_{M_x}(x)}{p_{M_{x|y}}(x|y)p_{M_y}(y)} dxdy \tag{2}$$

In this special case, the BYY learning is called Bayesian-Kullback YING-YANG (BKYY) learning. If there are a number of Ying-Yang systems with different architectures and the same or similar degree of *harmony*, we choose the one with the simplest complexity, as will be addressed later in Sec.2.3. The BYY theory provides a theoretical guide for a number of existing major learning models in parameter learning, regularization, structural scale or complexity selection, architecture design and data smoothing.

2.2 BKYY DIMENSION REDUCTION SYSTEM AND ARCHITECTURE DESIGN

This paper only concentrates on one special case of the BKYY system that maps input $x \in \Re^d$ into $y=[y^{(1)}, ..., y^{(k)}]^T$ with a reduced dimension $k < d$ such that a best reconstruction of x can be made from this y. For the purpose of 2D or 3D visualizations, we can simply fix $k = 2$ or 3. For general applications, we need to determine an appropriate dimension k.

In this system, the specification of $p_{M_x}(x)$ is usually straight-forward. Given a training set $D_x = \{x_i\}_{i=1}^N$, we simply let $p_{M_x}(x)$ fixed to a kernel estimate [Devroye 1987]:

$$p_h(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(x - x_i) \tag{3}$$

with a prefixed kernel function $K_h(x)$ and a smoothing parameter h. A special case that we often consider is that $h \to 0$ and thus $K_h(x - x_i) \to \delta(x - x_i)$.

We design $p_{M_y}(y)$ according to two different types of dimension reduction purposes:

(a) The system maps x into one of $n_{y|x}$ clusters or densities in a space of lower dimension. That is, the samples of y are assumed to spread in multiple modes, and thus parameterized by a finite mixture:

$$p_{M_y}(y) = p(y|\theta_k) = \sum_{j=1}^{n_y} \alpha_j p(y|\theta_j), \quad \alpha_j > 0, \quad \sum_{j=1}^{n_y} \alpha_j = 1 \tag{4}$$

In this way, the tasks of data dimension reduction and unsupervised classification are combined.

(b) The system reduces x to y, which consists of k independent components. That is, y is derived from a family of densities with independent components:

$$p_{M_y}(y) = p(y|\theta_k) = \prod_{j=1}^{k} p(y^{(j)}|\xi_j) \tag{5}$$

This case is actually a special case of BYY Dependency Reduction (BYY-DR) system and theory [Xu 1998] with $k < d$.

The specification of $p_{M_{x|y}}(x|y)$ is made by generating $x \in \Re^d$, $d > k$ from y via the Ying passage, which consists of $n_{x|y}$ linear or nonlinear channels, as follows:

$$x = g(y, A_j) + e_{x|y}^{(j)}, \quad j=1,..., n_{x|y}, \quad e_{x|y}^{(j)} \text{ from } p(e_{x|y}^{(j)}|\phi_j) \tag{6}$$

which is described by the finite mixture:

$$p_{M_{x|y}}(x|y) = \sum_{j=1}^{n_{x|y}} \gamma_j p(x - g(y, A_j)|\phi_j), \quad \gamma_j > 0, \quad \sum_{j=1}^{n_{x|y}} \gamma_j = 1 \tag{7}$$

From the fact that

$$\int p_{M_{x|y}}(x|y) p_{M_y}(y) dy = p_{M_2}(x) = \sum_{j=1}^{n_{x|y}} \gamma_j p(x|j), \quad p(x|j) = \int p(x - g(y, A_j)|\phi_j) p_{M_y}(y) dy \tag{8}$$

we see that the design of eq.(7) is equivalent to using a finite mixture of $p(x|j)$, $j=1, ..., n_{x|y}$ for modeling the marginal density $p(x)$.

Similarly, the specification of $p_{M_{y|x}}(y|x)$ is made by inverting x back to the original lower dimension y, via the Yang passage

$$y = f(x, U_j) + e_{y|x}^{(j)}, \quad j=1, ..., n_{y|x}, \quad e_{y|x}^{(j)} \text{ from } p(e_{y|x}^{(j)}|\psi_j) \tag{9}$$

that is also described by a finite mixture:

$$p_{M_{y|x}}(y|x) = \sum_{j=1}^{n_{y|x}} \beta_j p(y - f(x, U_j)|\psi_j), \quad \beta_j > 0, \quad \sum_{j=1}^{n_{y|x}} \beta_j = 1 \tag{10}$$

This design can be justified by the fact that it is equivalent to using the following finite mixture $p_{M_1}(y)$ for approximating $p_{M_y}(y)$:

$$\int p_{M_{y|x}}(y|x) p_h(x) dx = p_{M_1}(y) = \sum_{j=1}^{n_{y|x}} \beta_j p(y|j), \quad p(y|j) = \int p(y - f(x, U_j)|\psi_j) p_h(x) dx \tag{11}$$

A BKYY system with the above architecture design is called *BKYY Dimension Reduction* system.

2.3 BKYY DIMENSION REDUCTION AND DETERMINATION (BKYY-DRD) THEORY

Putting the four components $p_{M_{y|x}}(y|x)$, $p_{M_x}(x)$, $p_{M_{x|y}}(x|y)$ and $p_{M_y}(y)$ into eq.(2), we produce a BKYY-DRD theory, which consists of four parts, as follows:

(1) First, with k, $N=\{n_{x|y}, n_{y|x}, n_y\}$ fixed, we determine $\Theta_k=\{\beta_j,\gamma_j,\theta_k,\phi_j,\psi_j,A_j,U_j\}$ by

$$\Theta_k^* = \arg\min_{\Theta_k} KL(\Theta_k, N)\ ,\ KL \text{ given by eq.(2)}, \tag{12}$$

which is called *parameter learning* and can be implemented by an iterative *Alternative Minimization*:

$$\begin{aligned} &Step1: Fix M_2 = M_2^{old}, get M_1^{new} = \arg\min_{M_1} KL_{M_1,M_2} \\ &Step2: Fix M_1 = M_1^{old}, get M_2^{new} = \arg\min_{M_2} KL_{M_1,M_2} \end{aligned} \tag{13}$$

which guarantees to reduce KL_{M_1,M_2} until it converges to a local minimum at Θ_k^*.

(2) Second, with $N=\{n_{x|y}, n_{y|x}, n_y\}$ fixed, we determine the dimension k by

$$k^* = \min_k \mathrm{K}\ ,\quad \mathrm{K}=\{j: J_1(j)=\min_k J_1(k)\ ,\quad J_1(k)=KL(\Theta_k^*, \mathrm{N}) \tag{14}$$

That is, we determine the smallest k that minimizes $J_1(k)$. In other words, we select the simplest structural scale from multiple choices. It has been shown that, for a Ying-Yang pair with incremental architecture, $J_1(k) > J_1(k^o)$ for $k < k^o$ and $J_1^o(k)=J_1^o(k^o)$ for $k \geq k^o$, where k^o is the correct value for k.

For a set $D_x=\{x_t\}_{t=1}^N$ of finite samples, $J_1(k)$ may still slowly decrease even after $k \geq k^o$. Two types of solutions are suggested for detecting k^o in such cases. One selects k^o to be the point at which $J_1(k)$-$J_1(k+1)$ drops suddenly. The other way is to modify $J_1(k)$ into a new version $J_2(k)$ such that $J_2(k)$ decreases as k increases and reaches its minimum at k^o and then increases as k increases.

From eq.(2), we can decompose $J_1(k)$ into

$$J_1(k)=J_2(k)+H_{y|x}(k)\ ,\qquad H_{y|x}(k)=\int_{x,y} p_{M_{y|x}^*}(y|x)p_h(x)\ln p_{M_{y|x}^*}(y|x)dxdy \tag{15}$$

It is interesting to observe that $-H_{y|x}(k)$ is actually the average entropy of the Yang passage. For finite samples, the uncertainty of the Yang passage by $p_{M_{y|x}^*}(y|x)$ will increase when $k \geq k^o$, and thus the negative entropy $H_{y|x}(k)$ will decrease as k increases. Therefore, the removal of $H_{y|x}(k)$ from $J_1(k)$ will let $J_1(k)$-$H_{y|x}(k)$=$J_2(k)$ increase as k increases when $k \geq k^o$. Based on this, we propose the following alternative criterion:

$$k^* = \min_k J_2(k)\ ,\qquad J_2(k)=J_1(k)-\int_{x,y} p_{M_{y|x}^*}(y|x)p_h(x)\ln p_{M_{y|x}^*}(y|x)dxdy \tag{16}$$

for detecting k^o as the minimum point of $J_2(k)$, where $p_{M_{y|x}^*}(y|x)$ is obtained by minimizing $KL(\Theta_k, N)$. More generally, a similar result can be obtained by a family of $J_2(k)$ as

$$J_2(k)=J_1(k)-\gamma_r\int_{x,y} p_{M_{y|x}^*}(y|x)p_h(x)\ln p_{M_{y|x}^*}(y|x)dxdy\ ,\qquad 0\leq\gamma_r. \tag{17}$$

For the purpose of detecting the number of gaussians in a gaussian mixture, $J_2(k)$ given by eq.(16) are shown to work well both theoretically and experimentally. The criterion $J(k,\Theta_k)$ given by eq.(11) in Xu [1996] is actually a special case of $J_2(k)$ given by eq.(17) for a gaussian mixture with $\gamma_r=0.5$, which have been shown to work well experimentally.

(3) *Third*, we perform structural scale selection by $N^* = \min_N J(N)$ with

$$J(N)=\begin{cases} KL(\theta_k^*, N), & Corresponding\ to\ J_1, \\ KL(\theta_k^*, N)-\gamma_r\int_{x,y} p_{M_{y|x}^*}(y|x)p_h(x)\ln p_{M_{y|x}^*}(y|x)dxdy, 0\leq\gamma_r, & Corresponding\ to\ J_2. \end{cases} \tag{18}$$

(4) *Finally*, after learning, we map x back to y either stochastically according to $p_{M_{y|x}}(y|x)$ by eq.(10) or deterministically by taking the regression $E(y|x)=\int y p_{M_{y|x}}(y|x)dy$.

3 GAUSSIAN DESIGN, FACTOR ANALYSIS, PCA AND NONLINEAR PCA

3.1 GAUSSIAN MIXTURE BASED BKYY-DRD

We will consider specific cases of BKYY, based on gaussian mixtures:

$$p_{M_y}(y) = p(y \mid \theta_k) = \sum_{j=1}^{n_y} \alpha_j G(y, m_y^{(j)}, \lambda_y^{(j)}),$$

$$p_{M_{x|y}}(x \mid y) = \sum_{j=1}^{n_{x|y}} \gamma_j G(x, g(A_j y), \Sigma_{x|y}^{(j)}), \qquad p_{M_{y|x}}(y \mid x) = \sum_{j=1}^{n_{y|x}} \beta_j G(y, f(U_j x), \Sigma_{y|x}^{(j)}) \tag{19}$$

where $G(z, m, S)$ is a gaussian with mean vector m and co-variance matrix S, and $g(r)$, $f(r)$ are pre-specified nonlinear functions, e.g., a sigmoid function or even a complicated function implemented by a feed-forward network, with the degenerated case being the linear function $g(r) = r, f(r) = r$. Moreover, we have

$$g(Ay) = [g(a_1^T y), ..., g(a_n^T y)]^T, \; A^T = [a_1, ..., a_n],$$
$$f(Ux) = [f(u_1^T x), ..., f(u_k^T x)]^T, \; U^T = [u_1, ..., u_k], \tag{20}$$

which are usually called post-nonlinear mappings, because a univariate nonlinear function is imposed on the output of the linear mappings.

This system uses a mixture of post-nonlinear channels $g(A_j y)$ to generate x:

$$p_{M_2}(x) = \sum_{j=1}^{n_{x|y}} \gamma_j p(x \mid j), \; p(x \mid j) = \int G(x, g(A_j y), \Sigma_{x|y}^{(j)}) p_{M_y}(y) dy. \tag{21}$$

Also, this system uses a mixture of post-nonlinear channels $f(U_j x)$ to invert x back into one of $n_{y|x}$ distributions $p(y \mid j)$, $j = 1, ..., n_{y|x}$ in the dimension-reduced space:

$$p_{M_1}(y) = \sum_{j=1}^{n_{y|x}} \beta_j p(y \mid j), \; p(y \mid j) = \int G(y, f(U_j x), \Sigma_{y|x}^{(j)}) p_h(x) dx, \tag{22}$$

such that the tasks of data dimension reduction and unsupervised classification are combined.

With eq.(19), we get the following regressions in terms of linear combinations:

$$E(y \mid x) = \sum_{j=1}^{n_{y|x}} \beta_j f(U_j x), \; E(x \mid y) = \sum_{j=1}^{n_{x|y}} \gamma_j g(A_j y). \tag{23}$$

Through the use of random sampling techniques, we propose the following adaptive algorithm for parameter learning:

Step 1: Pick an integer j among $\{1, 2, ..., n_{y|x}\}$ according to the probabilities $\{\beta_1, ..., \beta_{n_{y|x}}\}$, and take a random sample ε_j from $G(\varepsilon_j, 0, \Sigma_{y|x}^{(j)old})$. Then let $y = f(U_j^{old} x) + \varepsilon_j$ for the current x.

Step 2:

(a) Let $h_y^{(j)} = \dfrac{\alpha_j^{old} G(y, m_y^{(j)old}, \lambda_y^{(j)old})}{\sum_{j=1}^{n_y} \alpha_j^{old} G(y, m_y^{(j)old}, \lambda_y^{(j)old})}$, update

$$\alpha_j = (1-\eta)\alpha_j^{old} + \eta h_y^{(j)}, \qquad m_y^{(j)} = m_y^{(j)old} + \eta \frac{h_y^{(j)}}{\alpha_j^{old}} (y - m_y^{(j)old}),$$

$$\lambda_y^{(j)} = (1-\eta)\lambda_y^{(j)old} + \eta \frac{h_y^{(j)}}{\alpha_j^{old}} (y - m_y^{(j)old})(y - m_y^{(j)old})^T.$$

(b) Let $h_{x|y}^{(j)} = \dfrac{\gamma_j^{old} G(x, g(A^{old} y), \Sigma_{x|y}^{(j)old})}{\sum_{j=1}^{n_{x|y}} \gamma_j^{old} G(x, g(A^{old} y), \Sigma_{x|y}^{(j)old})}$ and $\eta_g = \dfrac{dg(r)}{dr}\Big|_{r = A^{old} y}$, update

$$\gamma_j = (1-\eta)\gamma_j^{old} + \eta h_{x|y}^{(j)}, \qquad A_j = A_j^{old} + \eta \eta_g \frac{h_{x|y}^{(j)}}{\gamma_j^{old}} (\Sigma_{x|y}^{(j)old})^{-1} \left[x - g(A_j^{old} y)\right] y^T,$$

$$\Sigma_{x|y}^{(j)} = (1-\eta)\Sigma_{x|y}^{(j)old} + \eta\frac{h_{x|y}^{(j)}}{\gamma_j^{old}}\left[x - g(A_j^{old}y)\right]\left[x - g(A_j^{old}y)\right]^T,$$

(c) Let $h_{y|x}^{(j)} = \dfrac{\beta_j^{old}G(y, f(U_j^{old}x), \Sigma_{y|x}^{(j)old})}{\sum_{j=1}^{n_{y|x}}\beta_j^{old}G(y, f(U_j^{old}x), \Sigma_{y|x}^{(j)old})}$ and $\eta_f = \dfrac{df(r)}{dr}|_{r=U_j^{old}x}$, update

$$\beta_j = (1-\eta)\beta_j^{old} + \eta h_{y|x}^{(j)}, \quad U_j = U_j^{old} + \eta\eta_f\frac{h_{y|x}^{(j)}}{\beta_j^{old}}(\Sigma_{y|x}^{(j)old})^{-1}\left[y - f(U_j^{old}x)\right]x^T,$$

$$\Sigma_{y|x}^{(j)} = (1-\eta)\Sigma_{y|x}^{(j)old} + \eta\frac{h_{y|x}^{(j)}}{\beta_j^{old}}\left[y - f(U_j^{old}x)\right]\left[y - f(U_j^{old}x)\right]^T. \quad (24)$$

The above algorithm applies to the linear case of $g(r) = r, f(r) = r$ by simply inserting them in each place that they appear together with imposing $\eta_g = 1$, $\eta_f = 1$.

After applying the above parameter learning algorithm, we can also use eq.(14) and eq.(16) to determine the dimension k, as well as use eq.(18) for structure scale selection. In the next section, we will show that even the simplest case of eq.(19) with $n_y = 1$, $n_{y|x} = 1$ and $n_{x|y} = 1$ not only includes three existing dimension reduction methods as special cases, but also provides several interesting results.

3.2 Linear Reduction, Factor Analysis and Factor Number Selection

We begin by modeling a general linear dimension reduction problem. We consider an example where $y \in \Re^k$ comes from a gaussian distribution $p_{M_y}(y) = G(y,0,\lambda_y)$ with $E(y)=0$ and $\lambda_y = diag[\lambda_1^y,...,\lambda_k^y]$, $\lambda_j^y > 0$ and data $x \in \Re^n$, $n > k$ is generated from

$$x = Ay + e_x, \quad e_x \text{ is gaussian with } Ee_x = 0,\ E[e_x e_x^T] = \Sigma_{x|y},\ E[e_x y^T] = 0. \quad (25)$$

According to the derivation of Item 1 in Sec.4 later, without losing any generality, we can equivalently consider a simplified problem:

$$x = AD_k y + e_x, \text{ with } G(e_x,0,\Sigma_{x|y}) \text{ and } G(y,0,I_k),\ E[e_x y^T] = 0,\ A^TA = I_k,\ D_k = \text{diag}[d_1,...,d_k] > 0 \quad (26)$$

which actually defines a special architecture design as follows:

$$p_{M_y}(y) = G(y,0,I_k), \quad p_{M_{x|y}}(x \mid y) = G(x, AD_k y, \Sigma_{x|y}), \quad A^TA = I_k$$
$$p_{M_{y|x}}(y \mid x) = G(y, Ux, \Sigma_{y|x}), \quad p_{M_x}(x) = p_h(x) \text{ given by eq.(3)} \quad (27)$$

which is a simplest special case of eq.(19) with $g(r) = r, f(r) = r$ at $n_y = 1$, $n_{y|x} = 1$ and $n_{x|y} = 1$.

Putting the design into eq.(2), $\min_{\{U,\Sigma_{y|x}\}} KL(M_1, M_2)$ will result in

$$p(x,\Theta_k) = \int_y G(x, AD_k y, \Sigma_{x|y})G(y,0,I_k)dy = G(x,0,\Sigma_x),\ p_{M_{y|x}}(y \mid x) = \frac{G(x, AD_k y, \Sigma_{x|y})G(y,0,I_k)}{p(x,\Theta_k)}$$

or $U = D_k A^T\Sigma_x^{-1}$, $\Sigma_{y|x} = I_k - D_k A^T\Sigma_x^{-1}AD_k$, $\Sigma_x = \Sigma_{x|y} + AD_k^2A^T$. (28)

Thus, the minimization of $KL(M_1, M_2)$ becomes the following minimization:

$$\min_{\Theta_k} KL(\Theta_k), \quad \Theta_k = \{\Sigma_{x|y}, A\}, \quad KL(\Theta_k) = \int_x p_h(x)\ln\frac{p_h(x)}{p(x,\Theta_k)}dx. \quad (29)$$

which, as $h \to 0$, is equivalent to

$$\max_{\{\Sigma_{x|y},A,s.t.A^TA=I\}} L(\Theta_k), \quad L(\Theta_k) = \frac{1}{N}\sum_{i=1}^{N}\ln p(x_i,\Theta_k), \quad \text{with } p(x,\Theta_k) \text{ given by eq.(28).} \quad (30)$$

In other words, the special case is actually equivalent to conventional factor analysis, which has been widely studied in the literature of statistics [Sharma 1995].

This provides us with two benefits. The first is that we can now use the iterative *Alternative Minimization* procedure of eq.(13) to obtain the following new batch algorithm for implementing conventional factor analysis:

Step 1: Fix $A = A^{old}$, $D_k = D_k^{old}$, $\Sigma_x = \Sigma_x^{old}$, get $p_{M_{y|x}}(y \mid x) = G(y, Ux, \Sigma_{y|x})$, with

$U = D_k A^T\Sigma_x^{-1}$, $\Sigma_{y|x} = I_k - D_k A^T\Sigma_x^{-1}AD_k$;

Step 2: $\Sigma_{x|y} = \frac{1}{N}\sum_{i=1}^{N}\int_y p_{M_{y|x}}(y \mid x_i)(x_i - AD_k y)(x_i - AD_k y)^T dy = \frac{1}{N}\sum_{i=1}^{N}(x_i - AD_k Ux_i)(x_i - AD_k Ux_i)^T$,

Let $S = \frac{1}{N}\sum_{i=1}^{N}\Sigma_x^{-1} x_i \Sigma_{x|y}^{-1} x_i^T \Sigma_x^{-1}$ and solve the eigen-equation $SA = AD_k$ such that the column of A are the k eigenvectors of S that correspond to the first k eigenvalues.

Then , get $\Sigma_x = \Sigma_{x|y} + AD_k^2 A^T$, and let $A^{old} = A, D_k^{old} = D_k, \Sigma_x^{old} = \Sigma_x$. (31)

In addition, the special case of eq.(24) with $g(r) = r$, $f(r) = r$ at $n_y = 1$, $n_{y|x} = 1$ and $n_{x|y} = 1$ also provides us with an adaptive algorithm for performing factor analysis:

Step 1: $\Sigma_{y|x} = I_k - D_k A^T \Sigma_x^{-1} AD_k$, $U = D_k A^T \Sigma_x^{-1}$,

Take a sample x, then a random sample ε from $G(\varepsilon, 0, \Sigma_{y|x})$, let $y = Ux + \varepsilon$;

Step 2: $D_k = D_k^{old} + \eta A^T \Sigma_{x|y}^{-1\,old}(x - AD_k y)y^T$, $A = A^{old} + \eta(I - AA^T)\Sigma_{x|y}^{-1}(x - AD_k y)y^T D_k$,

$\Sigma_{x|y} = \Sigma_{x|y}^{old} + \eta(x - AD_k y)(x - AD_k y)^T$, $\Sigma_x = \Sigma_{x|y} + AD_k^2 A^T$, $U = D_k A^T \Sigma_x^{-1}$,

$A^{old} = A$, $D_k^{old} = D_k$, $\Sigma_{x|y}^{old} = \Sigma_{x|y}$. (32)

where $(I - AA^T)\Sigma_{x|y}^{-1}(x - AD_k y)y^T D_k$ is the gradient descent direction of $(x - AD_k y)\Sigma_{x|y}^{-1}(x - AD_k y)$ constrained on $A^T A = I_k$.

Both of the above algorithms should converge to a local maximum of $L(\Theta_k)$ given by eq.(30).

The second benfit is that we can now use $J_1(k)$ by eq.(14) or $J_2(k)$ from eq.(16) or eq.(17) to detect the unknown number of k factors. Actually, after ignoring some irrelevant constants, they can be simplified to:

$$J_1(k) = 0.5\left\{\ln\left|\Sigma_x^*\right| + Tr\left|\Sigma_x^{*-1} S_x\right|\right\}, \quad S_x = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T,$$

$$J_2(k) = \begin{cases} J_1(k) + 0.5(\ln\left|\Sigma_{y|x}^*\right| + k), & by\ eq.(16) \\ J_1(k) + 0.5\gamma_r(\ln\left|\Sigma_{y|x}^*\right| + k), 0 \le \gamma_r, & by\ eq.(17) \end{cases} \quad (33)$$

where '*' indicates the values obtained after convergence by eq.(31) or eq.(32).

3.3 PCA AND SUBSPACE DIMENSION DETERMINATION

We now consider a special case of eq.(27) with $\Sigma_{x|y} = \sigma_{x|y}^2 I_d$, where I_d is a d x d identity matrix. In this case, eq.(28) becomes

$$U = D_k A^T \Sigma_x^{-1}, \quad \Sigma_x = Adiag\left[(\sigma_{x|y}^2 I_k + D_k^2), \sigma_{x|y}^2 I_{d-k}\right]A^T,$$
$$\Sigma_{y|x} = I_k - D_k A^T \Sigma_x^{-1} AD_k = I_k - D_k^2(\sigma_{x|y}^2 I_k + D_k^2)^{-1}, \quad (34)$$

and thus eq.(29) and eq.(30) become

$\min_{\Theta_k} J(\Theta_k), J(\Theta_k) = \ln\left|\Sigma_x\right| + Tr\left[\Sigma_x^{-1} S_x\right]$, $\Sigma_x = Adiag\left[(\sigma_{x|y}^2 I_k + D_k^2), \sigma_{x|y}^2 I_{d-k}\right]A^T$, S_x given by eq.(31),

$$\ln\left|\Sigma_x\right| = \ln\left|\sigma_{x|y}^2 I_k + D_k^2\right| + (d-k)\ln\sigma_{x|y}^2 = \sum_{j=1}^{k}\ln(\sigma_{x|y}^2 + d_j^2) + (d-k)\ln\sigma_{x|y}^2,$$
$$Tr\left[\Sigma_x^{-1} S_x\right] = Tr\left\{Adiag[(\sigma_{x|y}^2 I_k + D_k^2), \sigma_{x|y}^2 I_{d-k}]^{-1} A^T S_x\right\}. \quad (35)$$

As shown in Item 2 of Sec.4, *its solution is that A consists of the first k principal component vectors of S_x as its column vectors*. This A can be solved by using either a batch or an adaptive algorithm, such as proposed in Xu [1993].

From eq.(29), we also have

$$U = diag[1 - \frac{\sigma_{x|y}^2}{\lambda_1^x}, ..., 1 - \frac{\sigma_{x|y}^2}{\lambda_k^x}]A^T, \Sigma_{y|x} = I_k - D_k^2 diag(\sigma_{x|y}^2 I_k + D_k^2)^{-1} = \sigma_{x|y}^2 diag[\frac{1}{\lambda_1^x}, ..., \frac{1}{\lambda_k^x}]. \quad (36)$$

Therefore, we can see that the BYY DRD for linear dimension reduction is equivalent to a modified PCA. Interestingly, we can prove that $E\|y-\hat{y}\|^2 = E\|y-Ux\|^2$ is minimized at U by eq.(36) with the minimum $Tr[\Sigma_{y|x}] = \sigma^2_{x|y}\sum_{j=1}^{k}\frac{1}{\lambda^x_j}$. That is, the error $E\|y-\hat{y}\|^2$ provided by the original PCA mapping (i.e., $U=A^T$) is larger than that produced by eq.(36).

Similar to eq.(33), we can also detect the subspace dimension k by

$$J_1(k) = 0.5\left[\sum_{j=1}^{k}\ln\lambda^x_j + (d-k)\ln\sigma^2_{x|y}\right],\quad \sigma^2_{x|y} = \frac{1}{d-k}\sum_{j=k+1}^{d}\lambda^x_j,\ \text{with}\ \lambda^x_j - \sigma^2_{x|y} > 0,$$

$$J_2(k) = \begin{cases} 0.5(d\ln\sigma^2_{x|y} + k), & \text{byeq.(16)}, \\ 0.5\left[(1-\gamma_r)\sum_{j=1}^{k}\ln\lambda^x_j + (d-k+\gamma_r k)\ln\sigma^2_{x|y} + \gamma_r k\right], & 0\le\gamma_r,\ \text{byeq.(17)}. \end{cases} \tag{37}$$

The criteria can be justified by the following two theorems:

Theorem 1 *(1) $J_1(k)$ is monotonically non-increasing as k increases. If $\lambda^x_1 > \ldots > \lambda^x_d$, $J_1(k)$ is monotonically decreasing as k increases.*

(2) Assume that the correct dimension of the subspace is k^o, we have

$$\lambda^x_1 \ge \ldots \ge \lambda^x_{k^o} \ge \lambda^x_{k^o+1} = \ldots = \lambda^x_d, \text{with}\ \lambda^x_j = \begin{cases} d_j^2 + \sigma^2_{o,x|y}, & j \le k^o, \\ \sigma^2_{o,x|y}, & j > k^o. \end{cases} \tag{38}$$

Thus, $J_1(k)$ is monotonically non-increasing as k increases from k = 1 to k^o and reaches the minimum before or at k^o, and then remains at this minimum as k continues to increase. Particularly, in eq.(38) if $\lambda^x_1 > \ldots > \lambda^x_{k^o}$, $J_1(k)$ monotonically decreases as k increases from k = 1 to k^o and reaches the minimum at k^o, and then remains at this minimum as k continues to increase.

Theorem 2 *Let $r(k+1) = \frac{\lambda^x_{k+1}}{\sigma^2_{k+1,x|y}} \ge 1$ which is actually the ratio of the $k+1^{th}$ eigenvalue over the average of the rest d-k-1 smallest eigenvalues.*

(1) $J_2(k)$ with $\gamma_r = 1$ is monotonically decreasing as k increases for $k \le k^o$ and reaches the minimum at k^o if $r(k+1) > (d-k)e^{1/d} - d + k + 1$ for $k > k^o$, and then monotonically increases as k continues to increase if $r(k+1) < (d-k)e^{1/d} - d + k + 1$.

(2) For $k > k^o$ we have r(k+1) = 1 and $J_2(k)$ with $\gamma_r = 1$ always monotonically increases as k continues to increase (since $(d-k)e^{1/d} > d + k$).

3.4 NONLINEAR PCA AND LMSER

We consider another special case of eq.(19) at $n_y = 1$, $n_{y|x} = 1$ and $n_{x|y} = 1$ with linear $g(Ay) = Ay$ but $f(Ux)$ being still post-nonlinear. That is, we have

$$p_{M_y}(y) = G(y, m_y, \lambda_y),\quad p_{M_{x|y}}(x\,|\,y) = G(x, Ay, \Sigma_{x|y}),\quad p_{M_{y|x}}(y\,|\,x) = G(y, f(Ux), \Sigma_{y|x}). \tag{39}$$

According to the derivation of Item 3 later in Sec.4, from eq.(3) with h = 0 we find that the minimization of KL(M_1, M_2) with respect to $\Theta_k = \{m_y, \lambda_y, A, \Sigma_{x|y}, U, \Sigma_{y|x}\}$ will become $\min_{\Theta_k} J(\Theta_k, k)$ with

$$J(\Theta_k, k) = \ln|\Sigma_{x|y}| + \frac{1}{N}\sum_{i=1}^{N}\left\|f(Ux_i) - \frac{1}{N}\sum_{i=1}^{N}f(Ux_i)\right\|^2,\quad \Sigma_{x|y}\ \text{given by eq.(48)}. \tag{40}$$

Therefore, we see that this case is equivalent to minimizing the volume of the covariance of the reconstruction error x_i - $Af(Ux_i)$ and the variance of $f(Ux_i)$. When $\Sigma_{x|y} = \sigma^2 I_d$, we have

$$J(\Theta_k, k) = d\ln\sigma^2 + \frac{1}{N}\sum_{i=1}^{N}\left\|f(Ux_i) - \frac{1}{N}\sum_{i=1}^{N}f(Ux_i)\right\|^2,\quad \sigma^2 = \frac{1}{dN}\sum_{i=1}^{N}\|x_i - Af(Ux_i)\|^2 \tag{41}$$

The minimization of the first term is equivalent to the minimization of σ^2, which represents the *auto-association* network of Bourlard & Kamp [1988]. Moreover, if we constrain $A = U^T$, we get

$$\sigma^2 = \frac{1}{dN}\sum_{i=1}^{N}\left\|x_i - U^T f(Ux_i)\right\|^2 . \tag{42}$$

Minimizing eq.(42) is equivalent to the *least mean squared error reconstruction* (LMSER) learning principle proposed in Xu [1993], which derived both the batch and adaptive gradient algorithms, and described a *symmetry breaking* property of nonlinearity. LMSER learning and its adaptive gradient algorithm was directly adopted to implement *Independent Component Analysis* by Karhunen, & Joutsensalo [1994] under the name of *Nonlinear PCA*.

Here, using eq.(40) and eq.(41), we describe several additional benefits:

(1) The minimization of the variance given by the second term in eq.(41) provides a regularization tool for reducing the range or uncertainty of the hidden representation $f(Ux_i)$. That is, using eq.(41) and eq.(42), we can obtain regularized *auto-association* and LMSER.

(2) Eq.(40) provides a more generalized version for *auto-association* and LMSER such that a general type of reconstruction error can be considered.

(3) We can use $J_1(k)$ from eq.(14) or $J_2(k)$ from eq.(16) or eq.(17) to detect the unknown dimension k. In this case, after ignoring some irrelevant constants, from eq.(40) and eq.(41) we have:

$$J_1(k) = \begin{cases} J(\Theta_k^*, k) \text{ by eq.(40) with } \Sigma_{x|y} \text{ by eq.(48) or } \Sigma_{x|y} = \frac{1}{N}\sum_{i=1}^{N}\left[x_i - U^T f(Ux_i)\right]\left[x_i - U^T f(Ux_i)\right]^T, \\ J(\Theta_k^*, k) \text{ by eq.(41) or } \sigma^2 \text{ by eq.(42)}, \end{cases}$$

$$J_2(k) = J_1(k) - \gamma_r \int p_{M_{y|x}^*}(y|x)p_h(x)\ln p_{M_{y|x}^*}(y|x)dxdy = J_1(k) + 0.5\gamma_r(\ln\left|\Sigma_{y|x}^*\right| + k). \tag{43}$$

(4) We have the following adaptive algorithm for implementing the general problem of minimizing $\ln\left|\Sigma_{x|y}\right|$ in eq.(40):

Step 1: Take a sample x, then a random sample ε from $G(\varepsilon, 0, I_k)$, let $y = U^{old}x + \varepsilon$;

Step 2: $A = A^{old} + \eta\Sigma_{x|y}^{-1}\left[x - A^{old}f(U^{old}x)\right]f(U^{old}x)^T$,

$$U = U^{old} + \eta(A^{old})^T\Sigma_{x|y}^{-1}\left[x - A^{old}f(U^{old}x)\right]x^T,$$

$$\Sigma_{x|y} = \Sigma_{x|y}^{old} + \eta\left[\left[x - A^{old}f(U^{old}x)\right]\left[x - A^{old}f(U^{old}x)\right]^T\right],$$

or particularly for the case $A = U^T$, we have

$$U^T = U^{T\,old} + \eta\Sigma_{x|y}^{-1}\left[x - U^{T\,old}f(U^{old}x)\right]f(U^{old}x)^T,$$

$$U^{new} = U + \eta U\Sigma_{x|y}^{-1}\left[x - U^{T\,old}f(U^{old}x)\right]x^T,$$

$$\Sigma_{x|y} = \Sigma_{x|y}^{old} + \eta\left[\left[x - U^{T\,old}f(U^{old}x)\right]\left[x - U^{T\,old}f(U^{old}x)\right]^T\right],$$

$$U^{old} = U^{new}, \Sigma_{x|y}^{old} = \Sigma_{x|y}. \tag{44}$$

Particularly, when $\Sigma_{x|y} = \sigma^2 I_d$, we can omit its updating and get

Step 2: $A = A^{old} + \eta\left[x - A^{old}f(U^{old}x)\right]f(U^{old}x)^T$, $U = U^{old} + \eta(A^{old})^T\left[x - A^{old}f(U^{old}x)\right]x^T$,

or particularly for the case $A = U^T$, we have

$$U^T = U^{T\,old} + \eta\left[x - U^{T\,old}f(U^{old}x)\right]f(U^{old}x)^T,\ U^{new} = U(I + \eta\left[x - U^{T\,old}f(U^{old}x)\right]x^T).$$

4 MATHEMATICAL DERIVATION

Item 1 [*The derivation of eq.(26)*] Because of the indeterminacy that $x = Ay + e_x = A'By + e_x = A'y' + e_x$, for any covariance matrix λ_y we have $I = \lambda_y' = B\lambda_y B^T$. Thus, without losing generality, we can assume that $y \in \Re^k$ is from a gaussian distribution $G(y, 0, I_k)$. Moreover, for a full rank matrix A, we have $A = UI_A D_k V^T$, $U^T U = UU^T = I_d$,

$V^TV = VV^T = I_k$, D_k = diag[d_1, ..., d_k] and $Ay = UI_AD_kV^Ty = UI_AD_ky'$ with the covariance of y' being still $V^TI_kV = I_k$, where I_A is a d x k matrix with its elements being 1 for $i = j$ and being 0 for $i \neq j$.

Item 2 [*The derivation for the solution of eq. (35)*] The minimization of $J(\Theta_k)$ by eq.(35) with respect to A with $A^TA = I_k$ can be de-coupled into the problem of minimizing $Tr[\Sigma_x^{-1} S_x]$ with respect to A. According to Theorem 4 Brockett [1989], all its local minima are reached only when A consists of the eigenvectors of S_x as the column vectors. That is,

$$A^TS_xA = \lambda^x, \quad \lambda^x = diag[\lambda_1^x,...,\lambda_k^x].$$

$$Tr[\Sigma_x^{-1} S_x] = Tr[diag[(\sigma_{x|y}^2 I_k + D_k^2), \sigma_{x|y}^2 I_{d-k}]^{-1}\lambda^x] = \sum_{j=1}^{k}(\sigma_{x|y}^2 + \gamma_j^2)^{-1}\lambda_j^x + \frac{\sum_{j=k+1}^{d}\lambda_j^x}{\sigma_{x|y}^2},$$

$$J(\Theta_k) = \sum_{j=1}^{k}\ln(\sigma_{x|y}^2 + \gamma_j^2) + (d-k)\ln\sigma_{x|y}^2 + \sum_{j=1}^{k}(\sigma_{x|y}^2 + d_j^2)^{-1}\lambda_j^x + \frac{\sum_{j=k+1}^{d}\lambda_j^x}{\sigma_{x|y}^2}. \tag{45}$$

Moreover, from $\frac{dJ(\Theta_k)}{d\gamma_j^2} = 0$ we have $\lambda_j^x = \sigma_{x|y}^2 + \gamma_j^2$ and

$$J(\Theta_k) = \sum_{j=1}^{k}\ln\lambda_j^x + (d-k)\ln\sigma_{x|y}^2 + k + \frac{\sum_{j=k+1}^{d}\lambda_j^x}{\sigma_{x|y}^2},$$

From $\frac{dJ(\Theta_k)}{d\sigma_{x|y}^2} = 0$, we have $\sigma_{x|y}^2 = \frac{\sum_{j=k+1}^{d}\lambda_j^x}{d-k}$ and $J(\Theta_k) = \sum_{j=1}^{k}\ln\lambda_j^x + (d-k)\ln\sigma_{x|y}^2 + d$.

Therefore, our problem becomes one of finding a permutation π of the indices {1, 2, ..., k} that minimizes the following criterion:

$$J_1(k) = 0.5\left[\sum_{j=1}^{k}\ln\lambda_j^x + (d-k)\ln\sigma_{x|y}^2\right], \quad \sigma_{x|y}^2 = \frac{1}{d-k}\sum_{j=k+1}^{d}\lambda_j^x, \text{ with } \lambda_j^x - \sigma_{x|y}^2 > 0 \tag{46}$$

The global minimum is achieved when $\lambda_1^x,...,\lambda_k^x$ are the first k largest eigenvalues of S_x.

Item 3 [*The derivation of eq. (40)*] After neglecting the irrelevant term $\int p_h(x)\ln p_h(x)dx$, from eq.(3) with h = 0, the minimization of $KL(M_1,M_2)$ with respect to $\Theta_k = \{m_y, \lambda_y, A, \Sigma_{x|y}, U, \Sigma_{y|x}\}$ will become $\min_{\Theta_k} J(\Theta_k, k)$ with

$$J(\Theta_k, k) = \frac{1}{N}\sum_{i=1}^{N}\int G(y, f(Ux_i), \Sigma_{y|x})\ln\frac{G(y, f(Ux_i), \Sigma_{y|x})}{G(y, m_y, \lambda_y)}dy - \frac{1}{N}\sum_{i=1}^{N}\int G(y, f(Ux_i), \Sigma_{y|x})\ln G(x_i, Ay, \Sigma_{x|y})dy$$

Further ignoring some constant, we have

$$2J(\Theta_k, k) = -\ln\frac{|\Sigma_{y|x}|}{|\lambda_y||\Sigma_{x|y}|} - k + \frac{1}{N}\sum_{i=1}^{N}Tr\left[\lambda_y^{-1}\int G(y, f(Ux_i), \Sigma_{y|x})(y-m_y)(y-m_y)^T dy\right]$$

$$+\frac{1}{N}\sum_{i=1}^{N}Tr[\Sigma_{x|y}^{-1}\int G(y, f(Ux_i), \Sigma_{y|x})(x_i - Ay)(x_i - Ay)^T dy]$$

$$= -\ln\frac{|\Sigma_{y|x}|}{|\lambda_y||\Sigma_{x|y}|} - k + \frac{1}{N}\sum_{i=1}^{N}Tr\left[\lambda_y^{-1}\int G(y, f(Ux_i), \Sigma_{y|x})(yy^T - m_yy^T - ym_y^T + m_ym_y^T)dy\right]$$

$$+\frac{1}{N}\sum_{i=1}^{N}Tr[\Sigma_{x|y}^{-1}\int G(y, f(Ux_i), \Sigma_{y|x})(x_ix_i^T - Ayx_i^T - x_i(Ay)^T + Ayy^TA^T)dy]$$

$$= -\ln\frac{|\Sigma_{y|x}|}{|\lambda_y||\Sigma_{x|y}|} - k + Tr[\lambda_y^{-1}\Sigma_{y|x}] + \frac{1}{N}\sum_{i=1}^{N}Tr\left[f(Ux_i) - m_y\right]^T\lambda_y^{-1}\left[f(Ux_i) - m_y\right]$$

$$+\frac{1}{N}\sum_{i=1}^{N}[x_i - Af(Ux_i)]^T\Sigma_{x|y}^{-1}[x_i - Af(Ux_i)] \tag{47}$$

From $dJ(\Theta_k, k) = -Tr(\Sigma_{y|x}^{-1} d\Sigma_{y|x}) + Tr[\lambda_y^{-1} d\Sigma_{y|x}]$, we have $\Sigma_{y|x} = \lambda_y$, thus $Tr[\lambda_y^{-1} \Sigma_{y|x}] = k$. Also, we have

$$\Sigma_{x|y} = \frac{1}{N}\sum_{i=1}^{N}[x_i - Af(Ux_i)][x_i - Af(Ux_i)]^T \ , \quad m_y = \frac{1}{N}\sum_{i=1}^{N} f(Ux_i) \ . \tag{48}$$

Moreover, for the same argument as made in the above Item 1, without losing generality we can let $\lambda_y = I_k$. Therefore, by ignoring the constant scale 2 in eq.(47) and also a constant $Tr[I_d] = d$ we get eq.(40).

5 EXPERIMENTS

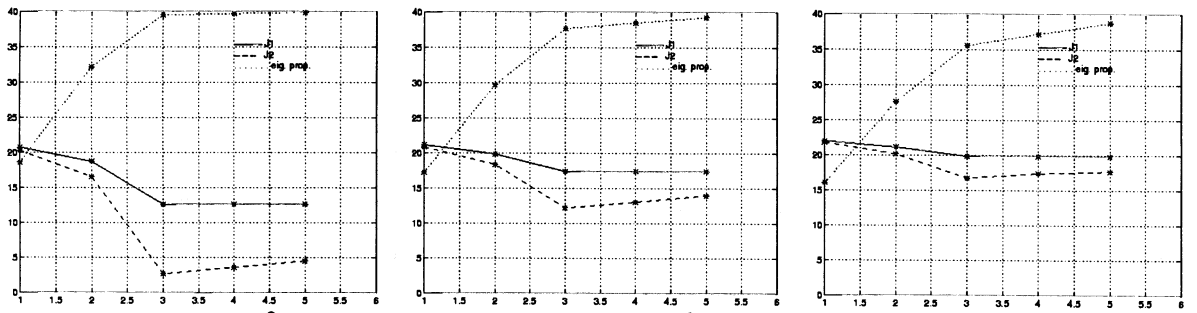
Here, we provide only some of the experimental results with regard to verifying the criteria $J_1(k)$ and $J_2(k)$ given in eq.(37) in comparison with a widely used conventional heuristic criterion, as follow:

$$\eta(k) = \sum_{j=1}^{k} \lambda_j^x \Big/ \sum_{j=1}^{d} \lambda_j^x > T_h \ , \tag{49}$$

with $0 < T_h < 1$ being a prespecified threshold Oja [1983]. Thus, the problem of selecting k is transformed to the problem of selecting T_h.

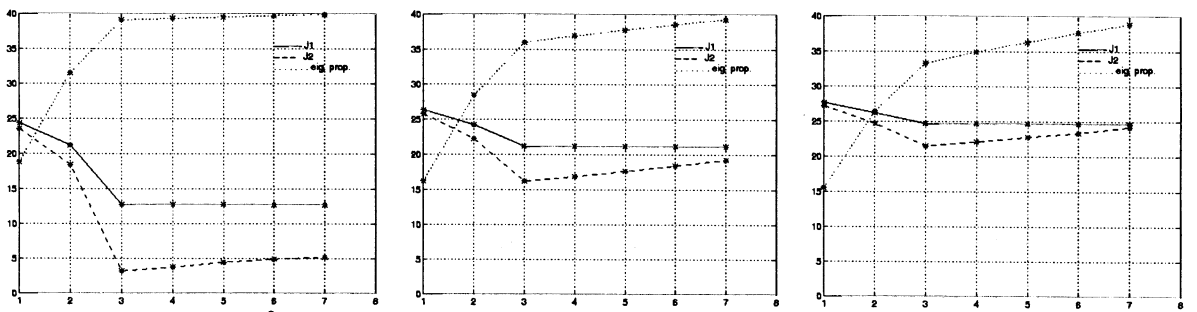
Samples points of $G(y, 0, I_3)$ are generated in $\Re^3$ and then mapped to x in $\Re^6$, $\Re^8$, $\Re^{10}$ by eq.(26), with $A^TA = I$ and $D_k^2 = diag[100,70,40]$ and noise added at weak, middle and strong levels with $\sigma_{x|y}^2 = 1, 5, 10$, respectively.

From Figures 2, 3 and 4, we see that both $J_1(k)$ and $J_2(k)$ monotonically decrease as k increases from $k = 1$ and reach corresponding minimum values at the correct dimension 3. Then $J_1(k)$ remains at its minimum value as k continues to increase, and $J_2(k)$ monotonically increases as k continues to increase. That is, Theorems 1 & 2 are verified. However, we see that the heuristic criterion $\eta(k)$ is always monotonically increasing. Although one may still detect the dimension by finding point on $\eta(k)$ where its rate-of-increase slows down, this is not as direct and robust as $J_1(k)$ and $J_2(k)$. Also, when noise is large, the detection of this point on $\eta(k)$ becomes more difficult as shown in Figures 2, 3 and 4.



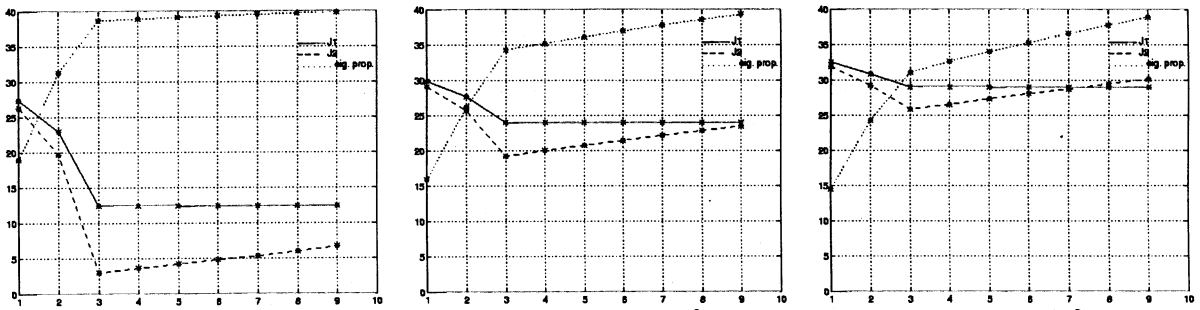
(a) noise variance $\sigma_{x|y}^2 = 1$ (b) noise variance $\sigma_{x|y}^2 = 5$ (c) noise variance $\sigma_{x|y}^2 = 10$

FIGURE 2. The resulted plots of J_1, J_2 and $\eta(k)$ on the data set that is mapped from $\Re^3$ to $\Re^6$, with $\eta(k)$ denoted by "eig.prop." (i.e., eigenvalues' proportion).



(a) noise variance $\sigma_{x|y}^2 = 1$ (b) noise variance $\sigma_{x|y}^2 = 5$ (c) noise variance $\sigma_{x|y}^2 = 10$

FIGURE 3. The plots of J_1, J_2 and $\eta(k)$ on the data set that is mapped from $\Re^3$ to $\Re^8$.



(a) noise variance $\sigma^2_{x|y} = 1$ (b) noise variance $\sigma^2_{x|y} = 5$ (c) noise variance $\sigma^2_{x|y} = 10$

FIGURE 4. The plots of J_1, J_2 and $\eta(k)$ on the data set that is mapped from $\Re^3$ to $\Re^{10}$.

6 CONCLUSIONS

A particular case of the Bayesian Ying-Yang (BYY) learning theory provides a general theory for developing various linear and nonlinear techniques for dimension reduction and determination (DRD). This theory not only includes factor analysis, PCA, *auto-association* networks, and LMSER-based nonlinear PCA as special cases, but also provides a number of new results, consisting of (a) new batch and adaptive algorithms for factor analysis, (b) criteria for determining the number of factors and the dimension of the PCA subspace, (c) a procedure for a specific Gaussian mixture based nonlinear BYY DRD, and (d) extensions for *auto-association* and LMSER-based nonlinear PCA.

ACKNOWLEGMENTS

The author would like to thank Mr. Wing-kai Lam for the help in preparing Figures 2 - 4. This work was supported by the HK RGC Earmarked Grant CUHK484/95E.

REFERENCES

Bourlard, H. and Y. Kamp [1989] "Auto-association by multilayer perceptron and singular value decomposition," *Biological Cybernetics*, Vol.*59*, pp.291-294.

Brockett, R.W. [1989] "Least Square matching problems," *Linear Algebra and Its Applications*, Vol.124, pp.761-777.

Devijver, P.A. and J. Kittler [1982] *Pattern Recognition: A statistical approach*, Prentice-Hall.

Karhunen, J. and JJ. Joutsensalo [1994] "Representation and separation of signals using nonlinear PCA type Learning," *Neural Networks*, Vol.7, pp.113-127.

Kohonen, T. [1995] *Self-Organizing Maps*, Springer-Verlag, Berlin.

Oja, E. [1983] *Subspace Methods of Pattern Recognition*, Research Studies Press, Letchworth, UK.

Oja. E.[1989] "Neural networks, principal components, and subspaces", *Int. J. Neural Systems*, Vol.1, pp.61-68.

Samon, John W., Jr. [1969] "A Nonlinear Mapping for Data Structure Analysis", *IEEE Trans. on Computers*, Vol.C-18, No.5.

Sharma, S. [1995] *Applied Multivariate Techniques*, John Wiley & Sons.

Tukey, John W. [1977] *Exploratory Data Analysis*, Addison-Wesley.

Xu, L. [1993] "Least mean square error reconstruction for self-organizing neural-nets," *Neural Networks*, Vol.6, pp.627-648.

Xu, L. [1994] "Theories for Unsupervised Learning: PCA and Its Nonlinear Extensions," Invited Talk, *Proc. of 1994 IEEE Intl. Conf. on Neural Networks*, pp.1253-1257.

Xu, L. [1995] "YING-YANG Machine: a Bayesian-Kullback scheme for unified learnings and new results on vector quantization," Keynote talk, *Proc. Int.l Conf. on Neural Information Processing* (ICONIP95), pp.977-988.

Xu, L. [1996] "How many clusters?: a YING-YANG machine-based theory for a classical open problem in pattern recognition," *Proc. of 1996 IEEE Intl. Conf. on Neural Networks*, pp.1546-1551.

Xu, L. [1997a] "Bayesian Ying-Yang System and Theory as A Unified Statistical Learning Approach: (I) Unsupervised and Semi-Unsupervised Learning," in *Brain-like Computing and Intelligent Information Systems*, Springer-Verlag, pp.241-274.

Xu, L. [1997b] "Bayesian Ying-Yang Machine, Clustering and Number of Clusters," *Pattern Recognition Letters*, Vol.18, No.11-13, pp.1167-1178.

Xu, L. [1998] "Bayesian Kullback Ying-Yang Dependence Reduction Theory," *Neurocomputing* (forthcoming).

Lei Xu is currently a professor with the Dept. of Computer Science and Engineering at the Chinese University Hong Kong, which he joined in 1993 as a senior lecturer. He has also been a professor at Peking University since 1992. During 1989-93, he worked as a postdoc or senior research associate at several universities in Finland, Canada and the USA, including Harvard and MIT. He is a past president of Asian-Pacific Neural Networks Assembly, an associate editor for several journals on neurocomputing, and has published over 180 papers. He serves on the editorial board of the Journal of Computational Intelligence in Finance, *and can be reached at the Dept. of Computer Science and Engineering, Chinese University of Hong Kong, Shatin, Hong Kong. Phone: 852 2609 8423, Fax: 852 2603 5024, Email: lxu@cse.cuhk.edu.hk.*

TABLE OF NOTATIONS

BYY- DRD	Bayesian Ying-Yang dimension reduction and determination
PCA	principal component analysis
LMSER	least mean square error reconstruction
$Tr[C]$	the trace of the square matrix C
diag$[d_1, ..., d_m]$	m x m diagonal matrix with its diagonal elements $d_1, ..., d_m$
I_m	m x m identity matrix
$x=[x_1, ..., x_d]^T \in \Re^d$	a data point in the original space of dimension d
$y=[y_1, ..., y_k]^T \in \Re^k$	a data point in the space of a reduced dimension $k < d$
Ux , $U^T = [u_1, ..., u_k]$	a liner dimension reduction mapping
$f(x, U)$	a nonlinear dimension reduction mapping in general case
$f(Ux)=[f(u_1^T x),..., f(u_k^T x)]^T$	post -nonlinear mapping with $f(r)$ being a univariate function
$M_1 = \{M_{y\|x}, M_x\}$ and $p_{M_1}(x, y)$	Yang machine and its representation in a joint distribution
$M_2 = \{M_{x\|y}, M_y\}$ and $p_{M_2}(x, y)$	Ying machine and its representation in a joint distribution
$y\|x$ and $p_{M_{y\|x}}(y\|x)$	the Yang passage $x \to y$ and its representation in a distribution
$x\|y$ and $p_{M_{x\|y}}(x\|y)$	the Ying passage $y \to x$ and its representation in a distribution
$p_{M_x}(x)$	the representation of the original space in a distribution
$K_h(x)$	a kernel function with smoothing parameter h
$p_h(x)$	a kernel estimation of $p_{M_x}(x)$, see eq.(3)
$p_{M_y}(y)$	the representation of the dimension reduced space in a distribution $p(y\|\theta_k)$
	a parametric representation of $p_{M_y}(y)$, e.g., see eq.(4)
α_j, θ_j	see eq.(4)
$x = g(y,A_j) + e_{x\|y}^{(j)}$	a Ying passage, e.g., see eq.(9)
γ_j, ϕ_j	see eq.(7)
$A^T = [a_1, ..., a_n]$	the parametric matrix in a backward mapping
$g(Ay)=[g(a_1^T y),..., g(a_n^T y)]^T$	a post -nonlinear backward mapping, $g(r)$ is a univariate function
$p_{M_2}(x)$	the marginal represented by Ying machine, see eq.(8), eq.(21)
$y = f(x,U_j) + e_{y\|x}^{(j)}$	a Yang passage, e.g., see eq.(6)
β_j, ψ_j	see eq.(10)
$p_{M_1}(y)$	the marginal represented by Yang machine, see eq.(11), eq.(22)
KL_{M_1,M_2}	the Kullback Divergence between Ying and Yang, see eq.(2)
$KL(\Theta_k, N)$	KL_{M_1,M_2} in its parametric representation, see eq.(12)
$J_1(k)$, $J_2(k)$	criteria for determining the dimension k, see eq.(14) and eq.(17)
$0 \le \gamma_r$	see eq.(17)
$\Theta_k^*, p_{M_{y\|x}^*}(y\|x)$	the specifications of Θ_k , $p_{M_{y\|x}}(y\|x)$ after parameter learning eq.(12)
k^*	the specifications of the dimension k by eq.(14) and eq.(17)
$G(z, m, S)$	a gaussian with mean m and co-variance S
$\lambda_y^{(j)}, \Sigma_{x\|y}^{(j)}, \Sigma_{y\|x}^{(j)}$	covariance matrices, see eq.(19)
$\eta > 0$	a pre-given learning stepsize
D_k	a diagonal matrix, see eq.(26)
$p(x,\Theta_k), \Sigma_k$	see eq.(28)
S_x	the sample covariance matrix, see eq.(33)
$\lambda_1^x,...,\lambda_k^x$	k eigenvalues of S_x